Can Frequentist Inferences Be Very Wrong?

A Conditional 'Yes'.*

by

David V. Hinkley

University of Minnesota

Technical Report No. 397

January 1982

This research supported in part by NSF Grant MCS-79-04558.

*This is an expanded version of a lecture given at the Conference on Scientific Inference, Data Analysis, and Robustness at Madison, Wisconsin in November 1981.

This research supported in part by NSF Grant MCS-79-04228.

*While in an exponential version of a Jacobite dyad of the conference on Potential Theory, Missouri in November 1981.

Can Frequentist Inferences Be Very Wrong?

A Conditional 'Yes'.

by

David V. Hinkley

"It is a capital mistake to theorize before one has data"

Sherlock Holmes, in Scandal in Bohemia

1. INTRODUCTION

The major operational difference between Bayesian and frequentist inferences is that in the latter one must choose a reference set for the sample, in order to obtain inferential probabilities. It is our thesis that in the matter of choosing a reference set, Sherlock Holmes was right, and that many frequentist inferences are inadequate because of erroneous choices made prior to the experiment.

It could not be argued with any conviction that the mathematization of statistics was other than very beneficial. The introduction of mathematics permitted logical development of the ideas of sufficiency, efficiency, hypothesis testing, design of experiments, quality control, multivariate data reduction and robustness, among others. Many of the current major advances in statistics rely heavily on sophisticated mathematics.

And yet, the formal structures of mathematical statistics seem to have a blind spot. Most of the mathematical development has to do with pre-data analysis: Is such-and-such likely to be a good procedure? How should we plan to do so-and-so? To answer such questions requires one to embed one's particular statistical problem a priori in a sample space with superimposed probability distribution over

potential realizations. The blind spot is the implicit assumption that pre-data probability calculations are relevant to post-data inferences. This blind spot is covered by Bayes's Theorem, which explicitly recognises the difference between pre-data and post-data contexts. Must the blind-spot remain in frequentist statistics? First we must recognize that it exists, and that is one purpose of the present paper: to show by example how naive frequentist answers can be misleading.

Some of our difficulties can be traced to the way in which we learn and teach probability and statistics. While we place great emphasis on the calculus of probability, we say little about the practical use of probability—other than to endorse the simple relative frequency interpretation. The following three probability statements suggest a variety of application that we would do well to understand and teach

P(coin will land head-up on a single flip) = ½

P(rain tomorrow in Madison) = ½

P(nuclear war in Europe before 1990) = ½.

It is evident that relative frequency in a real sequence of repeated experiments is simply not a rich enough interpretation of probability.

When we move from probability to theoretical statistics, we start by introducing such problems as "Let $X_1, \ldots, X_n$ be i.i.d. with p.d.f. $f_\theta(x)\ldots$" and "Let $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$." What meanings might these statements have? How do we determine that such statements are reasonable, even as approximations? I believe that there is far too little integration of inference with modelling and model diagnosis. For example, where do we read about inference subject to adequacy of model fit as judged by a goodness-of-fit test?

Once exact models are established, theoretical discussion focusses on exactness: unbiasedness, sufficiency, locally most powerful, inadmissibility, ancillarity, and so on. Of course these exact concepts are useful, in part because they prevent loose thought and ad hocery—but the concepts and exact properties should be used only as guides for careful

approximate analysis. Pedantic obsession with exactness can lead to absurdity, as a simple example will illustrate.

Example 1. Suppose that a population of N elements, with associated measurements $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_N, \tilde{y}_N)$, is sampled randomly n times with replacement. Let the generic sampled values be denoted by $X_i = \tilde{x}_{I_i}$ and $Y_i = \tilde{y}_{I_i}$, $i = 1, \ldots, n$; of course $\Pr(I_i = k) \equiv N^{-1}$. A graph of the data strongly suggests that $\tilde{y} \doteq \beta\tilde{x}$, and deviations from this approximate relationship seem reasonably random except for absolute magnitude. One might then propose the working model

$$Y_i = \beta X_i + \varepsilon_i \quad ,$$

with ε_i independent such that $E(\varepsilon_i | X_i = x) = 0$ and $\mathrm{Var}(\varepsilon_i | X_i = x) = \sigma^2(x)$, where the form of $\sigma^2(\cdot)$ is suggested by the sample. But now suppose that $\tilde{x}_1, \ldots, \tilde{x}_N$ are known and that they are distinct (not an unusual occurence). Then it cannot be true that $E(\varepsilon_i | X_i = x) = 0$, since $E(\varepsilon_i | X_i = \tilde{x}_j) = \tilde{y}_j - \beta\tilde{x}_j$, and further $\mathrm{Var}(\varepsilon_i | X_i = \tilde{x}_j) \equiv 0$. This truth is useless to practical analysis of the data. One reasonable mathematical way to proceed is by acknowledging that "$E(\varepsilon_i | X_i = x) = 0$" stands for "$|\mathrm{ave}(\tilde{y}_i - \beta\tilde{x}_i | x - \delta \leq \tilde{x}_i \leq x + \delta)| < \varepsilon$" for suitably small δ and ε—if a full-blown mathematical description is needed, which is doubtful.

This brief general discussion has raised questions about the difference between pre- and post-data probability calculations, the legitimacy of exact theory when integrated with practical application, and careful specification and understanding of what a statistical model means in practical terms. The purpose of the more detailed discussion which follows is to consider four topics where naive application of frequentist statistical theory can lead to incorrect or unhelpful inferences, whereas careful attention to the above questions can lead to sensible frquentist inferences.

The four topics to be discussed are: likelihood inference, inference from transformed data, randomization in design of experiments and surveys, and robust estimation.

One main thrust of the paper is that the general concept of ancillarity is an integral part of statistics. Beyond that, one might conclude that Bayesian inference provides a simple, direct way of obtaining sensible answers. This does not conflict with the view that frequency provides operational substance to many inferences. It is my curious belief that careful frequentist and careful Bayesian approaches can complement one another.

2. Likelihood Inference

The most clearly developed framework for inference principles and methods is that of stochastic models with a single unknown parameter. A conventional problem would treat observations $(x_1,\ldots,x_n) = \underset{\sim}{x}$ as a realization of random variables $(X_1,\ldots,X_n) = \underset{\sim}{X}$ whose joint distribution has density $f(\underset{\sim}{x}|\theta)$. Our discussion begins by reviewing what is known for the case of independent X_i, explained for simplicity under the assumption of homogeneity--that is, $f(\underset{\sim}{x}|\theta) = \Pi\, g(x_j|\theta)$. Part of our purpose is to expose an alternative large-sample approximate theory for likelihood inference, and to suggest by example how widely applicable that approximate theory may be.

There is agreement that for inference under a given model assumption, $\underset{\sim}{x}$ should first be reduced to the minimal sufficient statistic. In general this does not simplify matters much, so we must deal with approximate sufficiency. One program is to reduce

$$\underset{\sim}{X} \to (\hat{\theta},A_1,\ldots,A_p) = (\hat{\theta},\ \underset{\sim}{A})$$

where $\hat{\theta}$ is the MLE, and $\underset{\sim}{A}$ contains transformed characteristics of the likelihood shape. The larger is p, the less information is lost in the reduction in general. In particular, if $p = 1$ and if A_1 is the studentized form of the observed information $I = -\{d^2 \log f(\underset{\sim}{x}|\theta)/d\theta^2\}_{\theta=\hat{\theta}}$, then the information lost in reducing $\underset{\sim}{x} \to (\hat{\theta},I) \equiv (\hat{\theta},A_1)$ is $o(1)$ as $n \to \infty$. (Reduction $\underset{\sim}{x} \to \hat{\theta}$ does not have this property.) It is fairly clear that the rest of the likelihood shape, formalized in $(A_2,\ldots)$, can be ignored if the likelihood is very nearly normal in shape, which is often the case for

quite moderate sample size. It is also very clear from examples that the variation in spread of normal-shaped likelihoods can be appreciable, which is why reduction $\underset{\sim}{x} \to \hat{\theta}$ is often inadequate--as calculation of s.e. (I) would indicate.

Given appropriate (non-unique) definitions of $A_1, \ldots A_p$, one can show that

$$\begin{aligned} &\text{(i)} \quad I^{\frac{1}{2}}(\hat{\theta}-\theta) \approx N(0,1) \quad \text{given } \underset{\sim}{A}=\underset{\sim}{a} \,, \\ &\text{(ii)} \quad I/E(I) \approx N(1, n^{-1}c_\theta^2) \qquad c_\theta = 0(1) \,, \\ &\text{(iii)} \quad \underset{\sim}{A} \approx N_p(0,1). \end{aligned} \tag{2.1}$$

(For simplicity I write E(I) in place of $E\{-d^2 \log f(\underset{\sim}{x}|\theta)/d\theta^2\}_{\theta=\hat{\theta}}$, which it approximately equals.) These distributional approximations form the basis for an <u>approximate</u> analysis based on the <u>approximate</u> sufficient statistic $(\hat{\theta}, \underset{\sim}{A})$, and of course more refined approximations may sometimes be needed. If we were to reduce $\underset{\sim}{x} \to \hat{\theta}$, (i)-(iii) would not be directly relevant, but (i) and (ii) do imply

$$\text{(i')} \quad \{E(I)\}^{\frac{1}{2}} (\hat{\theta}-\theta) \approx N(0,1) \tag{2.2}$$

unconditionally, which is the classical first approximation result for the large-sample distribution of $\hat{\theta}$. The unusual derivation of (i') from (i) and (ii) will be important below.

The expressions (i) and (iii) give the (approximate) sufficient pivotal inference "$I^{\frac{1}{2}}(\hat{\theta}-\theta)$ is a standard normal random variable"--statement of the value $\underset{\sim}{a}$ of $\underset{\sim}{A}$ is uninformative. Why use the conditional statement (i) rather than (i')? To answer this in detail would be to repeat well-worn arguments for the conditionality principle, which we need not do; see Efron & Hinkley (1978). One brief answer will be given: if inference is contingent on adequacy of model fit, then the same (conditional) pivotal inference statement is valid, because the fit of the model is tested using $\underset{\sim}{A}$ (for example, the chi-squared statistic ΣA_j^2)--therefore the model adequacy restriction is covered by the conditioning on $\underset{\sim}{A}=\underset{\sim}{a}$. By contrast, the distributional approximation (i') may be recognizably invalid for subsets of the <u>a priori</u> sample space which are determined by restrictions on $\underset{\sim}{A}$.

It should be clear that (2.1) (i) may not be sufficiently accurate. Both Cox (1980) and Hinkley (1980) show that the likelihood itself can be integrated to give a more accurate approximation for the conditional distribution of $I^{\frac{1}{2}}(\hat{\theta}-\theta)$. This brings us tantalizingly close to the formal Bayesian posterior distribution. Notice that the Bayesian analysis requires no choice of pivot, which might prompt a frequentist to investigate the whole class of approximate conditional inferences based on pivots $Q(\hat{\theta},\theta,A_1,\ldots,A_p)$, rather than restricting attention to $I^{\frac{1}{2}}(\hat{\theta}-\theta)$. Of course the search for exact equivalence of conditional frequentist and Bayesian inferences is in principle hopeless, but it is of interest to study the approximate formal equivalence, since this gives added force to the Bayesian method.

What has been outlined here is a means of approximate analysis based on the actual information content of the data. The reader should consult Barndorff-Nielsen (1980), Amari (1982a,b), and other cited references, for theoretical details.

The question I should like to address now is the extent to which our discussion generalizes, beyond the case of independent sampling with fixed sample size. This draws us back to the relationships (2.1) (i) and (ii). No general theoretical account seems to be possible at this stage, so we shall look at two interesting examples. The first deals with a non-stationary process, and the second deals with random sample size.

Example 2. Autoregressive Process

Let $X_1,X_2,\ldots,X_n$ be modelled by the AR1 process

$$X_o = 0, \quad X_j = \theta X_{j-1} + \varepsilon_j , \quad j=1,\ldots,n ,$$

where $\varepsilon_1,\varepsilon_2,\ldots$ are independent N(0,1). The earlier discussion applies if $|\theta| < 1$, that is to say (2.1) and (2.2) hold, for large n, essentially because the process is ergodic. But for $|\theta| \geq 1$, when the process is not ergodic, it is well known that (i') fails. What seems not to be well known is

that (i) still holds for $|\theta| \geq 1$, but (ii) fails. This reinforces the idea that (i) is the fundamental normal approximation of likelihood inference.

The loglikelihood for θ based on $x_1,\ldots,x_n$ is

$$\ell(\theta) = \text{constant} + \theta\sum_1^n x_j x_{j-1} - \tfrac{1}{2}\theta^2 \sum_1^n x_{j-1}^2 ,$$

so that the sufficient statistic is the pair $(\hat\theta, I)$ where

$$\hat\theta = \sum_1^n x_j x_{j-1} / \sum_1^n x_{j-1}^2 \quad \text{and} \quad I = -\ddot{\ell}(\hat\theta) = \sum_1^n x_{j-1}^2 ;$$

the normalized likelihood is exactly $N(\hat\theta, I^{-1})$. For $|\theta| \geq 1$, the magnitude of variation in I prevents convergence of $I/E(I)$, and (i') fails: another unconditional limiting distribution obtains for $\{E(I)\}^{\frac{1}{2}}(\hat\theta-\theta)$. The insufficiency of $\hat\theta$ alone, and the huge asymptotic fluctuations of I, make the unconditional limiting result clearly irrelevant for finite sample analysis. The sufficient privotal result $I^{\frac{1}{2}}(\hat\theta-\theta) \approx N(0,1)$ has apparently been established for $\theta=1$ and large n by both D. Siegmund and T. Lai, and empirical evidence such as is described below suggests validity of the result for $|\theta| \geq 1$.

The unconditional theory for the case $\theta=1$ has been studied by Evans and Savin (1981), who enumerate the non-normal limiting distribution of $\{E(I)\}^{\frac{1}{2}}(\hat\theta-\theta)$; here $E(I) = \frac{1}{2}n(n-1)$ is the true expectation of I at $\theta=1$. Figure 1 shows (dotted curve) the induced approximate unconditional distribution of $\hat\theta-\theta$ for n=20. The figure also shows (dashed lines) the induced conditional normal approximations for $\hat\theta-\theta$ obtained by "undoing" $I^{\frac{1}{2}}(\hat\theta-\theta) \approx N(0,1)$ when $I=\frac{1}{4}$, 1 and 2 times E(I). (Simulation shows that when n=20, $\Pr\{I \leq \frac{1}{4}E(I)\} \approx \frac{1}{4}$ and $\Pr\{I \geq 2E(I)\} \approx 1/8$, so that Figure 1 encompasses a reasonable span for I.)

Evidently inference from the unconditional approximate distribution would be very misleading even if $I = E(I)$. For example, still with n=20, the unconditional result indicates that values of $\hat\theta$ smaller than $\theta - 0.4 = 0.6$ would be significant evidence that θ is less than 1, rather than equal to

Figure 1: Comparison of induced distributions for $\hat{\theta}-\theta$ in AR1 process with $x_o=0$ and n=20. Dotted line: unconditional asymptotic distribution when $\theta=1$. Dashed line: conditional normal approximations induced from N(0,1) approximation for $I^{\frac{1}{2}}(\hat{\theta}-\theta)$.

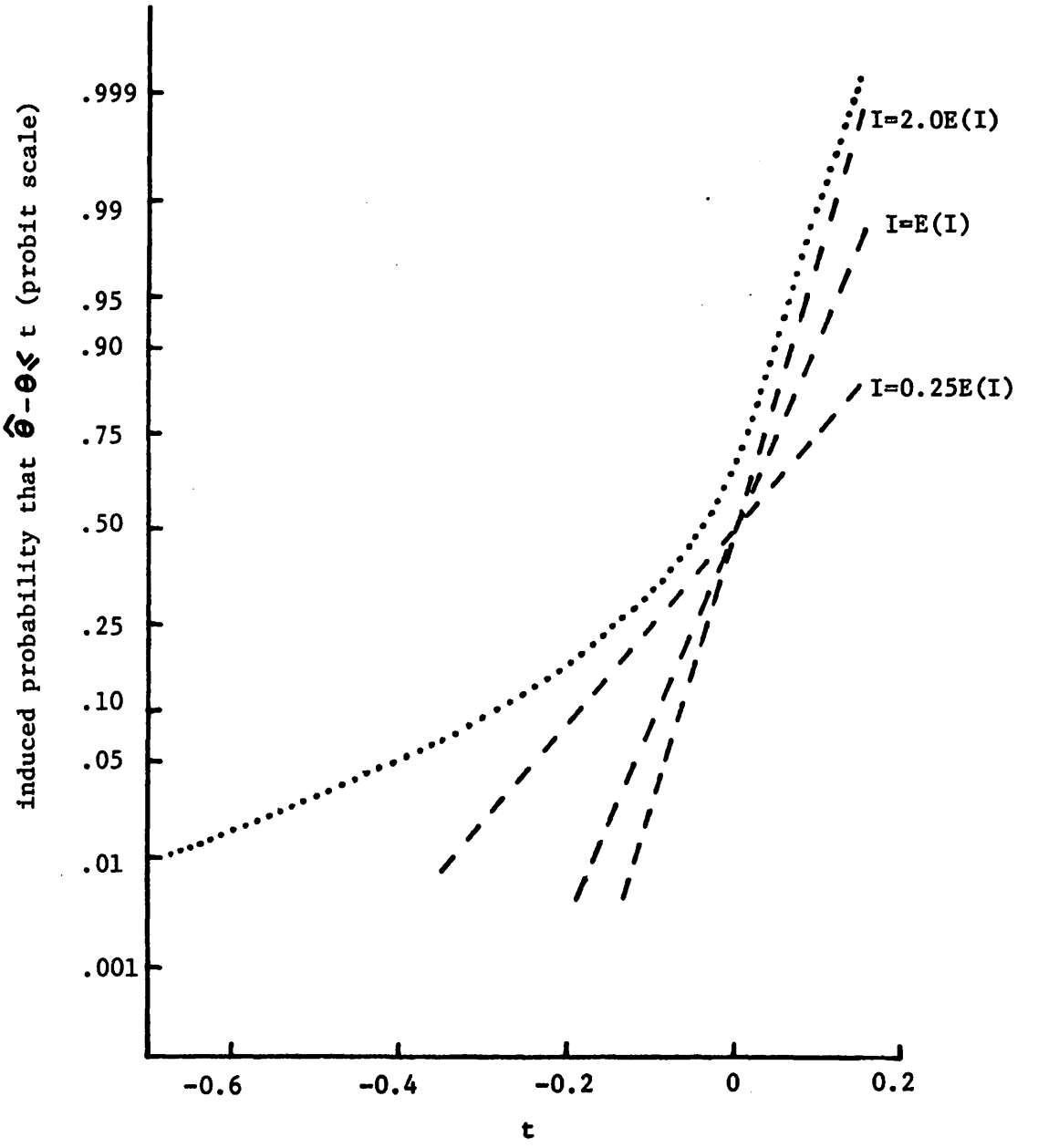
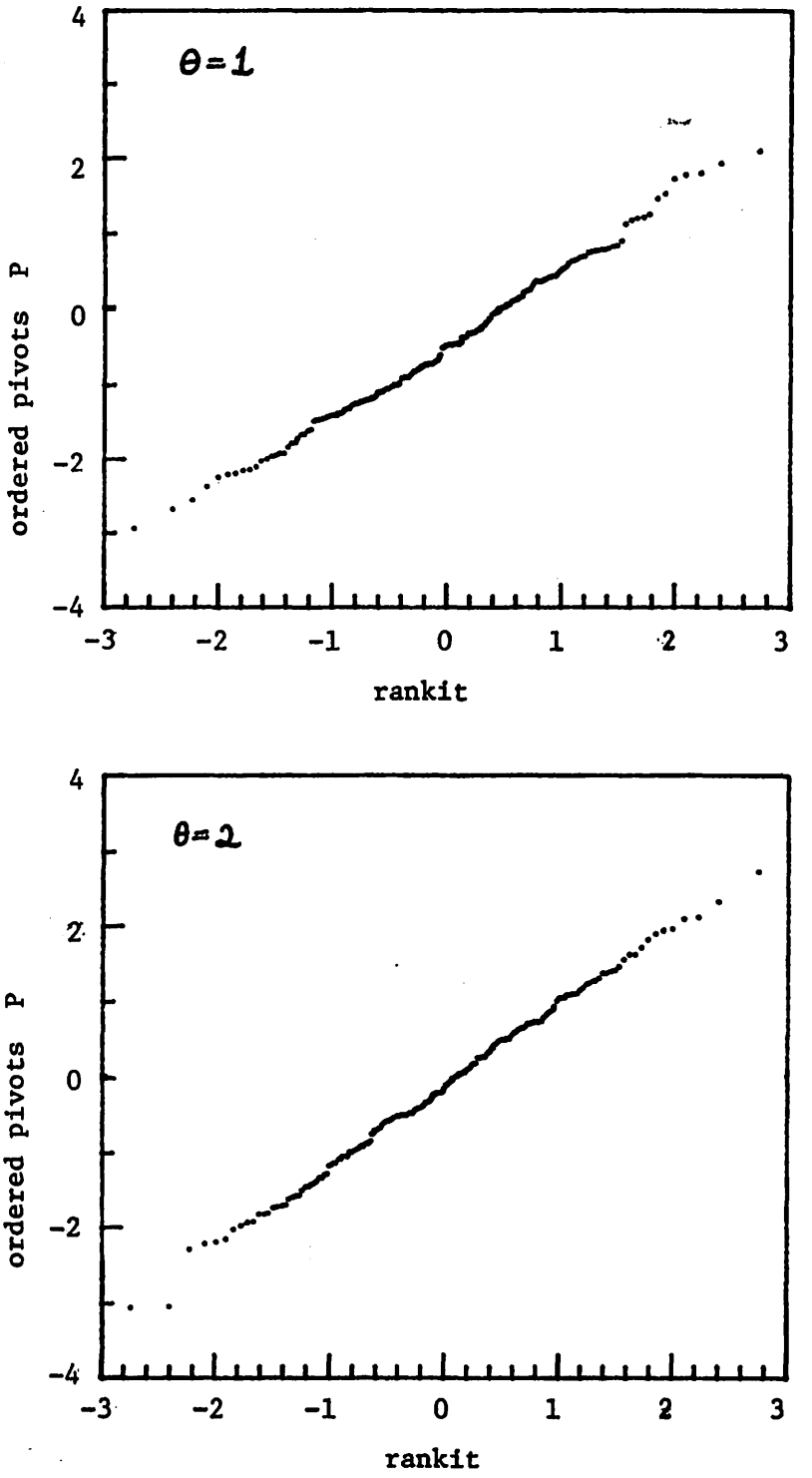


Figure 2: Rankit plots of 200 simulated values of pivot $P=I^{\frac{1}{2}}(\hat{\theta}-\theta)$ for AR1 process with $x_0=0$ and $n=20$.
Upper graph: $\theta=1$. Lower graph: $\theta=2$.



one, at the 5% level. But the conditional significance level at $\hat{\theta} = 0.6$ is 5% for $I = 1/10E(I)$, less than 1% for $I = \frac{1}{4}E(I)$ and essentially zero for $I \geq E(I)$.

Should econometricians test $\theta=1$ using the unconditional result, there will be a lot of random walks in econometric models!

Are the approximations in fact of reasonable accuracy? Yes and no. Figure 2 shows rankit plots of 200 simulated values of $I^{\frac{1}{2}}(\hat{\theta}-\theta)$ for the cases $\theta=1.0$ and 2.0 with $x_o=0$ and $n=20$. The plots for θ near 1 show a perceptible bias, which disappears for $\theta=2$. The folded distributions agree very well with theoretical approximation: in 10,000 cases with $\theta=1$ the frequencies with which $I(\hat{\theta}-\theta)^2$ exceeded the 2%, 5% and 10% points of χ^2_1 were respectively 1.98%, 4.99% and 9.98%.

Example 3. Sequential Sampling

Let $\{X(t):t \geq 0\}$ be a Wiener process with drift rate θ, i.e. increments $dX(t)$ are independent $N(\theta dt, dt)$. Anscombe (1957) considered this model under two sampling rules: (a) stop sampling at $T=t_o$, (b) stop sampling when first $X(t)=x_o$. In general, the stopping time T and $X \equiv X(T)$ are jointly sufficient for data-dependent stopping rules, and so a possibility of approximate ancillarity exists. But $\hat{\theta}=X/T$ is sufficient under either of rules (a) and (b). The normalized likelihood is always exactly $N(\hat{\theta}, I^{-1})$, $I \equiv T$. Under rule (a), $I^{\frac{1}{2}}(\hat{\theta}-\theta)$ is exactly $N(0,1)$, but the same is not so under rule (b). Thus, as Anscombe pointed out, at the level of exact inference it is not always possible to obtain frequency-based inference that agrees with the likelihood principle--a principle fundamental to Bayesian inference. Nevertheless, in the sample space in which samples are embedded to obtain frequency, $I^{\frac{1}{2}}(\hat{\theta}-\theta)$ is still approximately $N(0,1)$. The exact density of $Q = I^{\frac{1}{2}}(\hat{\theta}-\theta)$ under rule (b) is

$$p(q|x_o,\theta) = \{1 + \tfrac{1}{2}(\theta-\hat{\theta})\}^{-1} \phi(q) \quad ,$$

where $\hat{\theta}$ is a function of q. (Somewhat mischievously, the usable approximation $p(q|x_o,\hat{\theta})$ is exactly $\phi(q)$!) It would be of some interest to consider a general stopping rule with (T,X) both random.

A rather different sequential scheme was investigated by Grambsch (1980). She showed that for independent $X_1, X_2, \ldots, X_N$ with $N = \inf(n: I \geq i_o)$, still $I^{\frac{1}{2}}(\hat{\theta}-\theta) \approx N(0,1)$.

What these two examples suggest is that "$I^{\frac{1}{2}}(\hat{\theta}-\theta) \approx N(0,1)$" is a very general normal approximation, appropriate when the likelihood is approximately normal shaped. Or, put another way, the result of applying Bayes's theorem has approximate validity in a frequency sense if the sample sequence is embedded in a suitably small subset of the sample space--this "suitably small subset" being determined by an approximately ancillary statistic when possible. Much further work is needed before these vague ideas and suggestions can be turned into a concrete, reliable approximate theory--if this is indeed possible. What is not in doubt, however, is that unconditional asymptotic theory for $\hat{\theta}$ is insufficient both in the technical sense and in the practical sense.

3. Data-based Transformation: Box-Cox Revisited Again

A conventional mathematical formulation for linear regression on explanatory variables $\underset{\sim}{u}$ with random deviations is

$$Y_i = \theta^T u_i + \sigma\varepsilon_i \quad , \quad i=1,\ldots,n \quad , \tag{3.1}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent $N(0,1)$. In practise Y may be a derived measurement, e.g. logarithm or reciprocal of an actual measurement, and clearly some empirical judgement should be involved in such a particular choice of Y. Box & Cox (1964) proposed and studied a particularly useful class of derived measurements obtained from actual measurements X, namely $Y = (X^{\lambda}-1)/\lambda$. The data then provide evidence about those values of λ which are compatible with models of the type (3.1). The literal extension of (3.1) is

$$X_i(\lambda) = (X_i^{\lambda}-1)/\lambda = \theta^T u_i + \sigma\varepsilon_i \quad , \quad i=1,\ldots,n \; . \tag{3.2}$$

A maximum likelihood approach to model fitting is first to obtain $\hat{\lambda}$, and then to apply standard least squares to data $\{(u_i, x_i(\hat{\lambda})), \; i=1,\ldots,n\}$. But what of inference?

Bickel & Doksum (1981) show that under model (3.2), the distributions of $\hat{\theta}$ and $\hat{\sigma}^2$ can be very much more dispersed than if λ were known, and hence not estimated. For example, if $u \equiv 1$ and $\lambda = 0$, then $\sqrt{n}(\hat{\theta}-\theta) \approx N(0,\sigma^2+\zeta^2)$, $\zeta^2 = 1/6\sigma^2(1+\theta^4/\sigma^4)$--the extra term ζ^2 being "due to estimation of λ". To my chagrin, this particular result was first noted by myself (Hinkley, 1975).

Bickel & Doksum's remarkable observation is mathematically correct, but statistically incorrect: the variance inflation is not relevant to data analysis. This is an instance where the mathematical model has come adrift from its motivating moorings. A simple example will clarify the main point.

Suppose that we have two samples of positive measurements x--so that associated u vectors are (1,1) and (1,-1), say--and that we wish to compare the two sampled populations. The simplest type of comparison is via means, so we fit model (3.2) and find $\hat{\lambda}=0$. In all respects the two samples appear to suggest that (3.2) fits well. Whatever be the true population λ, our data model says that $\log_e X \sim N(\mu_i,\sigma^2)$ in sample i, i=1,2. For reasonably large n, then, our comparison is summarized by the following statement

$$\text{"}(\mu_1-\mu_2) - (\hat{\mu}_1-\hat{\mu}_2) \approx N(0,\sigma^2(1/n_1+1/n_2)) \qquad (3.3)$$

where μ_i is the mean of $\log_e X$ in population i"; n_1 and n_2 are sample sizes. In contrast to this statement is one based on Bickel & Doksum's results, which takes the form

$$\text{"}\theta_2-\hat{\theta}_2 \approx N(0,\sigma^2(1/n_1+1/n_2) + \zeta^2(1/n_1+1/n_2)) \qquad (3.4)$$

where θ_2 is ...". θ_2 is what? All that can be said is that θ_2 is the mean population contrast of some unknown transformation of X! The point is that (3.3) is a scientifically complete statement, whereas (3.4) is not. Further, (3.3) matches the interplay between data modelling and statistical inference--it is the same statement that would be made if we arrived at the log transformation by some other logical exploratory method.

The statement (3.3) deserves an extended discussion, which will be given elsewhere (Hinkley and Runger, 1982). Here only the key points will be mentioned. First, (3.3) is the inference in a context which does not make $\mu = \log_e X$ of a prior interest--that would involve a different analysis, clearly, since $\log_e X$ need not always appear to be normally distributed. Secondly, the probability result used in (3.3) needs justification. In the absence of the extended discussion, I will note that (3.3) is precisely the statement that you, the reader, would use if I presented you with the Normal-looking values of $\log_e X$ and asked for a statement about $\mu_1-\mu_2$. Thirdly, the difficulty can be ascribed to incompleteness of formulation: (3.2) should really be

$$X_i(\lambda) = \{\theta(\lambda)\}^T u_i + \sigma(\lambda)\varepsilon_i \quad , \quad i=1,\ldots,n \quad ,$$

where roughly speaking $\theta(m) = E(\hat{\theta}|\hat{\lambda}=m)$. (Box and Cox rescale $x(\lambda)$ to offset the incomparability of $\theta(\lambda)$ values as λ varies.) The Bickel & Doksum phenomenon of variance inflation is simply that

$$\mathrm{Var}(\hat{\theta}) = E\ \mathrm{Var}\{\hat{\theta}(\hat{\lambda})|\hat{\lambda}\} + \mathrm{Var}\ E\{\hat{\theta}(\hat{\lambda})|\hat{\lambda}\}$$

where the second term on the right is the inflation--which we necessarily discount when we fix our scale by $\hat{\lambda}_{obs}$ and make an inference about $\theta(\hat{\lambda}_{obs})$. Fourthly, there is the Bayesian approach. Note that the marginal posterior for θ will give, approximately, (3.4) and this is deemed to be irrelevant. What is required, if $\hat{\lambda}=0$ as before, is the posterior distribution for $\theta_2(0) = \mu_1-\mu_2$. This distribution can be derived, approximately, by writing $\theta(0) \approx \theta(\lambda)-\lambda\dot{\theta}_2(0)$. I leave the reader to follow this approach through to (3.3).

I have already mentioned that for a parameter of prior interest, such as $\mu_1-\mu_2$, the analysis would be different. So it would for a predictive statement about the observable X. For inference relative to a fixed <u>a priori</u> scale, the Bickel & Doksum results do not apply mathematically or statistically. For some relevant work see Carroll and Ruppert (1981).

The lessons to be learned here are (i) that care must be exercised in model statement (compare (3.2) and (3.3)), and (ii) that inference statements must be physically complete, i.e. well defined and relevant.

4. Randomized Designs and Their Analyses

Two subjects are of interest here, classical experimental design and sample survey analysis, the common element being randomization--in the first case used to allocate units to treatments, and in the second case used to select units for measurement. My main point is that the general concept of ancillarity plays an important role in both design and analysis, in the sense of pre- and post-design blocking.

Symbolic summaries of the randomized design problems would be as follows:

Experimental Design: Given a set of designs $\mathfrak{D} = \{d_1,\ldots,d_M\}$, apply the random design D where $Pr(D=d_j) \equiv M^{-1}$.

Sample Survey (with Replacement): Given population units $\tilde{u}_1,\ldots,\tilde{u}_N$, a design d is a point in $\{\tilde{u}_1,\ldots,\tilde{u}_N\}^n$. From the set $\mathfrak{D} = (d_1,\ldots,d_M)$, $M=N^n$, choose a random design D such that $Pr(D=d_i) \equiv M^{-1}$. (D involves individual selections $U_1,\ldots,U_n$ such that $Pr(U_i=u_j) \equiv N^{-1}$.)

It is sometimes argued that D is an ancillary statistic and hence that randomization theory of statistics is incompatible with the theory of conditioning on ancillary statistics. This interesting twist of logic belies the purpose of an ancillary statistic, which is to act as an indicator of appropriate reference sets. (Part of the absurdity of the exact logic was mentioned in Example 1.) One point that has been made is that randomization is often used to make normal-theory analysis valid (approximately), in which case D is ancillary by design. A second point is that the ancillary statistic should partition $\mathfrak{D}$ only as far as is statistically useful--i.e. as far as the relevant subsets of the a priori sample space. Two examples will clarify the situation.

Example 4. Knight's Move Latin Squares

In 1931, Tedin reported on a detailed analysis of the 5x5 Latin Squares as applied to some uniformity data (agricultural plot responses in the absence of treatment). These

particular data evidenced spatial correlation. The 5x5 Latin Squares $\mathfrak{D}$ can be subdivided into $\mathfrak{D}_1$ = Knight's Move Squares, $\mathfrak{D}_2$ = Diagonal Squares, $\mathfrak{D}_3$ = Other Squares. If we denote by σ_i the randomization standard error of a treatment contrast conditional on $D\varepsilon\mathfrak{D}_i$, then Tedin showed that $\sigma_1 < \sigma_3 < \sigma_2$ for his uniformity data. However, the estimated variance $\hat{\sigma}^2$ based on residual sum of squares behaves in contrary fashion, the restricted randomization means satisfying

$$E(\hat{\sigma}^2 | D\varepsilon\mathfrak{D}_1) > E(\hat{\sigma}^2 | D\varepsilon\mathfrak{D}_3) = \sigma_3^2 > E(\hat{\sigma}^2 | D\varepsilon\mathfrak{D}_2) .$$

The central equality validates normal-theory estimation of standard error for the restricted randomization over $\mathfrak{D}_3$. Thus the ancillary indicator of design subset can be used in conjunction with empirical data to facilitate design.

Of course if reliable estimates of σ_i were available, one could select D from $\mathfrak{D}$ without restriction and then use the ancillary subset indicator in the data analysis to obtain the relevant standard error of a contrast; $\mathfrak{D}_1$ would be a preferable restriction in such a situation. If a spatial correlation model accounted for the differences noted by Tedin, then presumably appropriate analyses under that model would distinguish between $\mathfrak{D}_1$, $\mathfrak{D}_2$ and $\mathfrak{D}_3$ on the basis of an (approximately) ancillary statistic.

Further discussion of this example may be found in Yates (1965).

Example 5. Ratio Estimation of a Population Total

The following situation seems to be common in some census problems. For a population of units $(\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_N)$, each of which possesses quantitative characteristics x and y, there is complete knowledge of $\tilde{x}_1, \ldots, \tilde{x}_N$. It is now desired to estimate $T = \sum_1^N \tilde{y}_i$, and the pairs (x,y) can be measured on a random sample of units. Thus we obtain a random sample $\{(x_1,y_1), \ldots, (x_n,y_n)\}$. The ratio estimator of T is

$$\hat{T} = (\sum_1^N \tilde{x}_i)\bar{y}/\bar{x} \quad , \qquad \bar{x} = n^{-1}\sum_1^n x_j \quad , \quad \bar{y} = n^{-1}\sum_1^n y_j \quad .$$

Several estimates V of $\mathrm{Var}(\hat{T})$ have been proposed, and some are justified on the grounds that their means over the full randomization distribution are nearly unbiased. Such justification is rightly challenged by Royall and Cumberland (1981), on the well-documented empirical evidence that often the values of $x_1,\ldots,x_n$ define relevant subsets of the full a priori sample space. Thus $\mathrm{Var}(\hat{T}|x_1,\ldots,x_n)$ and $E(V|x_1,\ldots,x_n)$--carefully interpreted--may be very unequal, as they both vary with the x_i. This could often be anticipated because in many cases $Y \approx \beta X + \sigma(x)e$ is a plausible model. Then the relationship between $(x_1,\ldots,x_n)$ and $(\tilde{x}_1,\ldots,\tilde{x}_N)$ can be used as an ancillary indicator in a conditional inference, based on estimate $\hat{T}$. In effect the indicator would act as a post-data stratification instrument.

In both examples it is suggested that a statistical indicator $a(D)$ be used to split $\mathfrak{D}$ into subsets, either for the purpose of design or for the purpose of analysis. In the latter case, the value of $a(D)$ will define a subset on which the inferential probability is defined, either by restricted randomization or by some plausible model that is validated by the randomization.

5. Robust Estimation

A great deal of effort has gone into the following problem: if $x_1,\ldots,x_n$ are independently distributed each with density of the form $f(x-\theta)$ where $f(y)$ is symmetric about $y=0$, what is a good estimate for θ if the form of f is unknown? Our language has been enriched by terms such as contamination, leverage and break-down point, and dozens of estimators have been studied intensively--both by brains and by computers. One thing seems to have been overlooked: how does one analyse a given data set?

A common approach to robustness theory might be described simply as follows: Consider a class $\mathfrak{T}$ of statistical functionals $t(\cdot)$ such that $t(\hat{F}_n)$ is an unbiased estimate of θ, $\hat{F}_n$ being the empirical distribution function. Then if attention can be restricted to a class $\mathfrak{F}$ of underlying

distribution functions F, choose the estimate $t(\hat{F}_n)$ to obtain $\min_{t\in\mathcal{T}} \max_{F\in\mathcal{F}} E\{t(F_n)-\theta\}^2$. Often F would be a neighborhood of a special distribution, such as the Normal.

An asymptotic theory for statistical functionals gives unconditional normal approximations for $t(\hat{F}_n)-\theta$. But this seems unsatisfactory on several grounds. First, parametric inference shows convincingly (Efron & Hinkley, 1978) that the relevant precision for $t(\hat{F}_n)$ should be tied to the residuals $\hat{e}_i = x_i - t(\hat{F}_n)$. Secondly, statisticians diagnose data with the aid of things such as normal plots, and can often gain information about which F is appropriate by using the residuals. Should we worry that F might be capable of producing very large deviations $x-\theta$, if in fact the given data yields the best normal plot we ever saw? Clearly not. Related to this is the more specific conjecture: that whatever F is, $\bar{X}$ is a pretty good estimate if a goodness-of-fit test of normality does not reject the normality hypothesis. This seems plausible because for any choice of t, $\bar{X} = t(\hat{F}_n) + d_t(\hat{e})$, whereas a goodness-of-fit statistic for a particular F must be of the form $g_{t,F}(\hat{e})$. The covariation of $(d_t(\hat{e}), g_{t,F}(\hat{e}))$ will keep $d_t(\hat{e})$ in check if $g_{t,F}(\hat{e})$ lies near its expected value under F.

Would it be reasonable to choose the estimator t and calculate its standard error as if $F=F_o$ when the observed data are compatible with F_o? In a general sense one would say "yes", because this is how applied statisticians behave when they first model their data. The theoretician would be doubtful. The key question might be posed in the following simple form: Suppose $\mathcal{F} = (F_1,\ldots,F_m)$, that

$$n\mathrm{Var}\{t(\hat{F}_n)\,|\,F_i,\hat{e}\} \approx \sigma_i^2(\hat{e}) \quad ,$$

and that $\hat{F}_{\mathcal{F}}$ is the F_j which is closest to $\hat{F}_n$ (in an appropriate sense). Is it then true that

$$n\mathrm{Var}\{t(\hat{F}_n)\,|\,F_i,\hat{F}_{\mathcal{F}}=F_k,\hat{e}\} \approx \sigma_k^2(\hat{e})?$$

(The same question is of interest for unconditional variances.) If true, this would parallel a Bayesian analysis for relatively uninformative priors:

$$p(\theta|x_1,\ldots,x_n)=C\sum_{F\in\mathcal{F}} p(\theta|x_1,\ldots,x_n,F)p(F|x_1,\ldots,x_n)p_{PRIOR}(F)$$

$$\approx p(\theta|t(\hat{F}_n),\hat{e},\hat{F}_{\mathcal{F}}),$$

which is approximately a $N(0, n^{-1}\sigma_k^2(\hat{e}))$ distribution for $\theta-t(\hat{F}_n)$ when $\hat{F}_{\mathcal{F}}=\hat{F}_k$.

Of course a Bayesian analysis has the general advantage of responding to specific features of the data. What I am suggesting is that careful attention to robust inference might reveal a sensible, responsive frequency theory which essentially justifies the results of a Bayesian analysis. The unconditional distribution theory prevalent in robustness literature is fine for choosing estimates which are generally good, but not for analysis of a particular data set.

REFERENCES

AMARI, S. (1982a). Differential geometry of curved exponential families--curvatures and information loss. Ann. Statist., 10 (June issue).

AMARI, S. (1982b). Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika, 69 (to appear).

ANSCOMBE, F.J. (1957). Dependence of the fiducial argument on the sampling rule. Biometrika, 44, 464-469.

BARNDORFF-NIELSEN, D. (1980). Conditionality resolutions. Biometrika, 67, 293-310.

BICKEL, P.J. and DOKSUM, K.A. (1981). An analysis of transformations revisited. J.Am.Statist.Assoc., 76, 296-311.

BOX, G.E.P. and COX, D.R. (1964). An analysis of transformations (with discussion). J.R.Statist.Soc., B26, 211-252.

CARROLL, R.J. and RUPPERT, D. (1981). On prediction and the power transformation family. Biometrika, 68, (to appear).

COX, D.R. (1980). Local ancillarity. Biometrika, 67, 279-286.

EFRON, B. and HINKLEY, D.V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. Biometrika, 65, 457-482.

EVANS, G.B.A. and SAVIN, N.E. (1981). The calculation of the limiting distribution of the least squares estimator of the parameter in a random walk model. Ann.Statist., 9, 1114-1118.

GRAMBSCH, P. (1980). Likelihood inference. Ph.D. Dissertation, University of Minnesota.

HINKLEY, D.V. (1975). On power transformations to symmetry. Biometrika, 62, 101-111.

HINKLEY, D.V. (1980). Likelihood as approximate pivotal distribution. Biometrika, 67, 287-292.

HINKLEY, D.V. and RUNGER, G. (1982). Analysis of Box-Cox transformed data. University of Minnesota School of Statistics Technical Report.

ROYALL, R.M. and CUMBERLAND, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance (with discussion). J.Ann.Statist.Assoc., 76, 66-88.

TEDIN, O. (1931). The influence of systematic plot arrangement upon the estimate of error in field experiments. J.Agric.Sci.Camb., 21, 191-208.

YATES, F. (1965). A fresh look at the basic principles of the design and analysis of experiments. Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 4, 777-790.

The author was partially supported by National Science Foundation Grant MCS-79-04558.

School of Statistics
University of Minnesota
Minneapolis, MN 55455